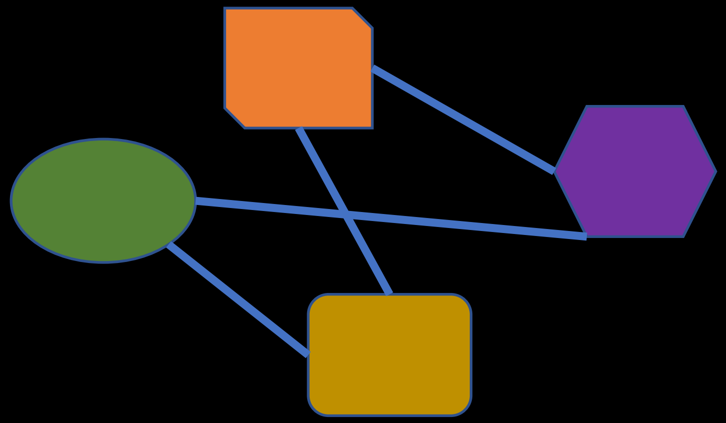
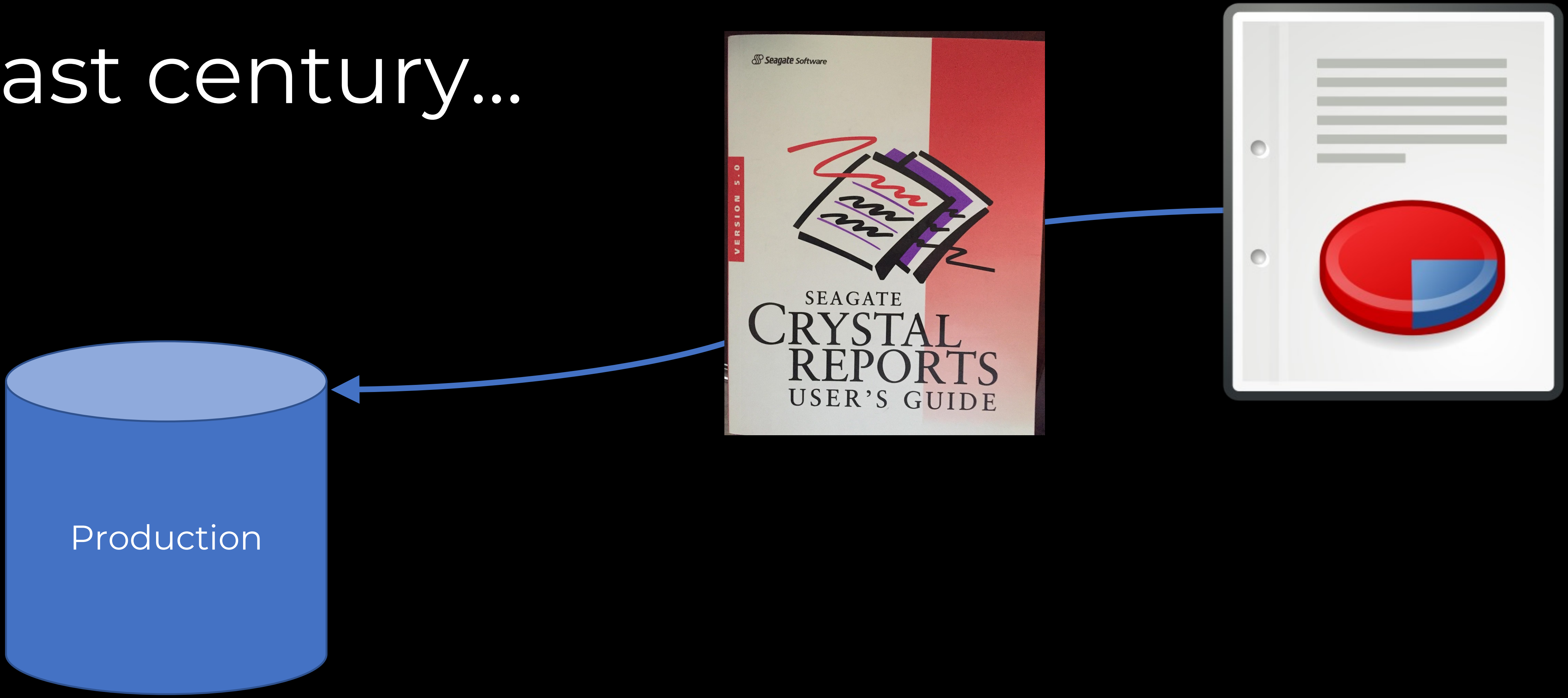


Data Mesh –
The emperor’s new clothes…

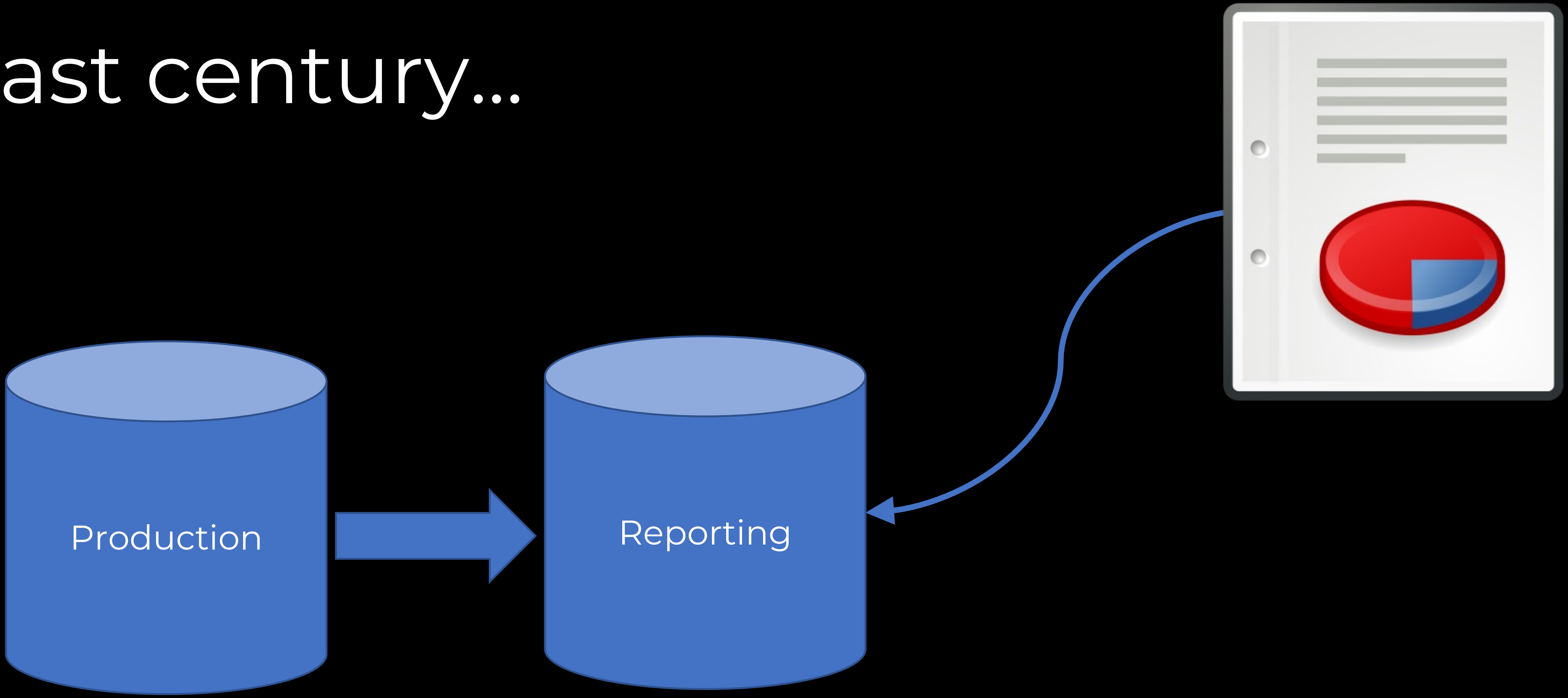
Rune Ovlien Rakeie, Johan Ludvig Brattås

Data Saturday Stockholm 2022

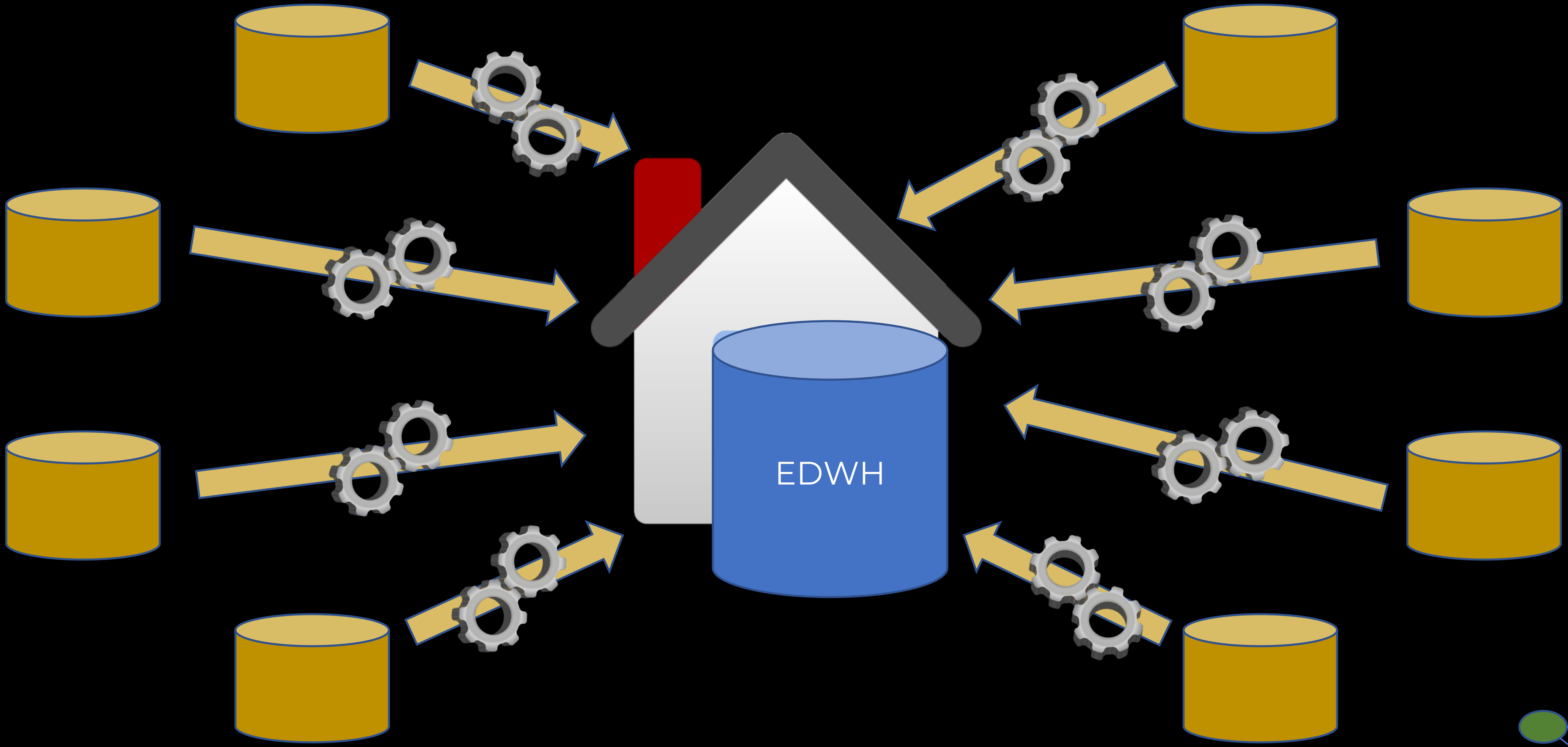
Last century...



Last century…

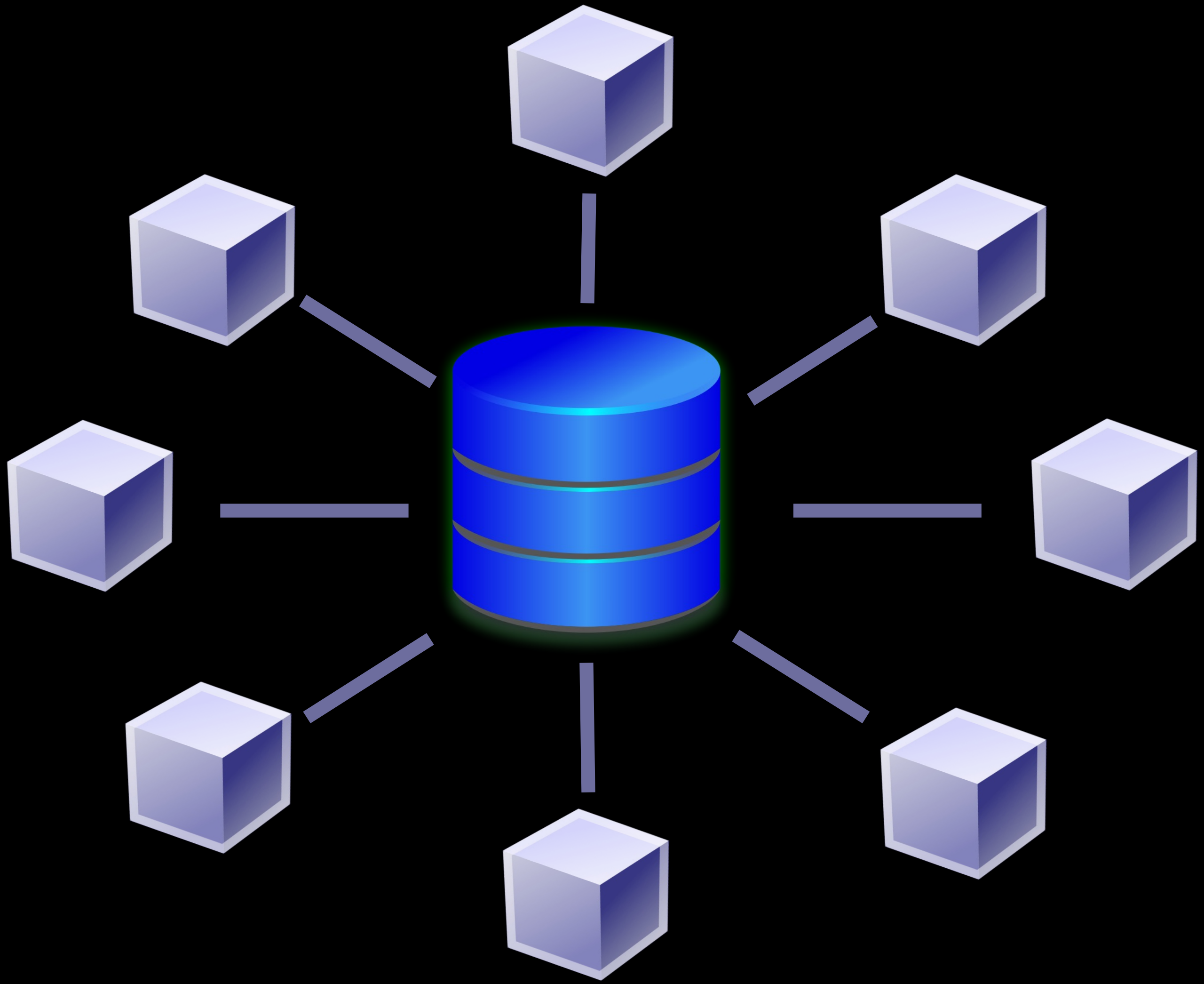


1. Generation
Enterprise data warehouse & BI



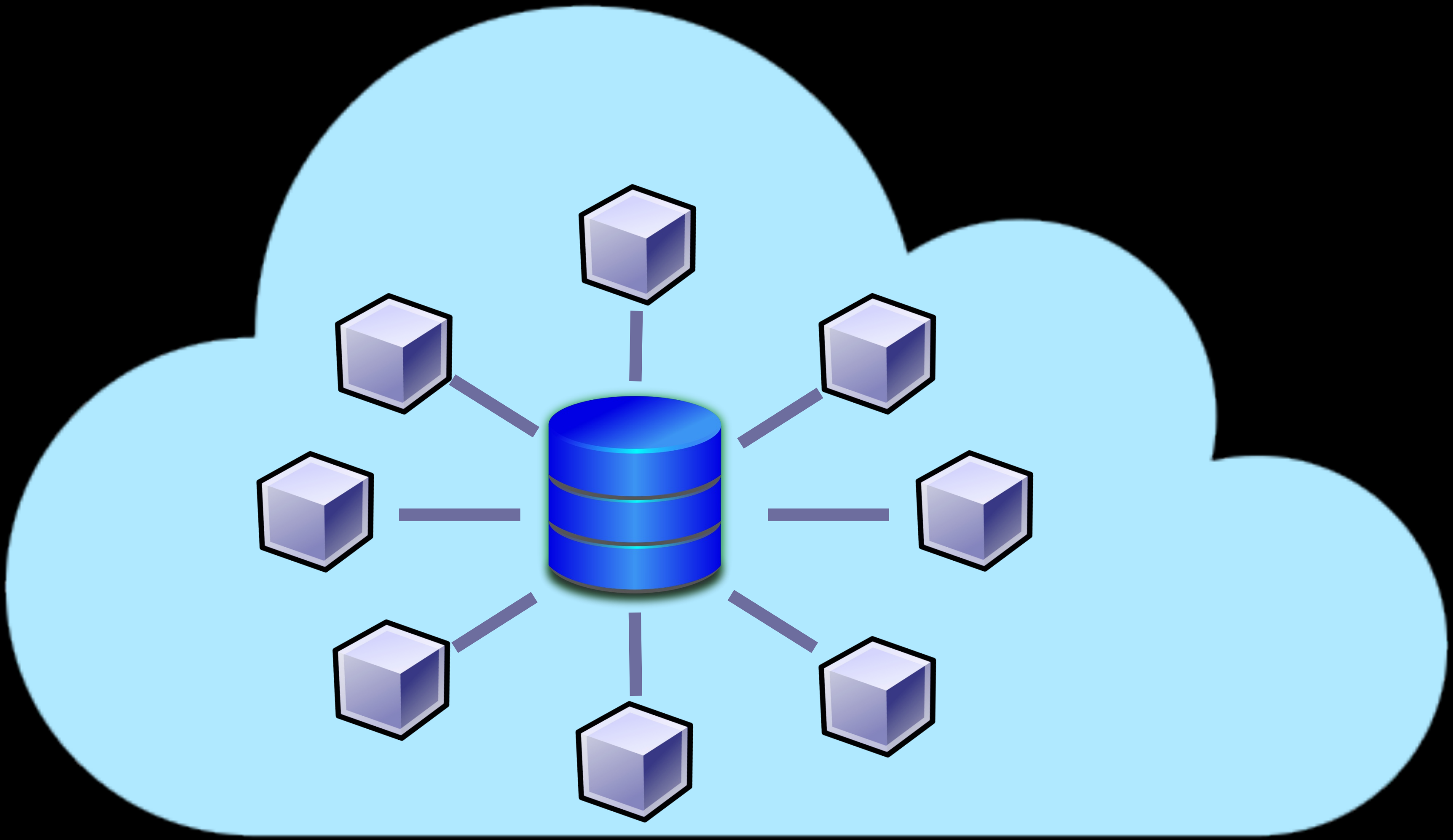
2. Generation

Big Data & data lake (→ swamp)



3. Generation

2. Gen + streaming data and cloud



Data Mesh origins

How to Move Beyond a Monolithic Data Lake to a Distributed Data Mesh

(*20 May 2019, martinfowler.com*)

Data Mesh Principles and Logical Architecture

(03 *December 2020, martinfowler.com*)

Zhamak Dehghani
Thoughtworks

Problems Data Mesh aim to solve

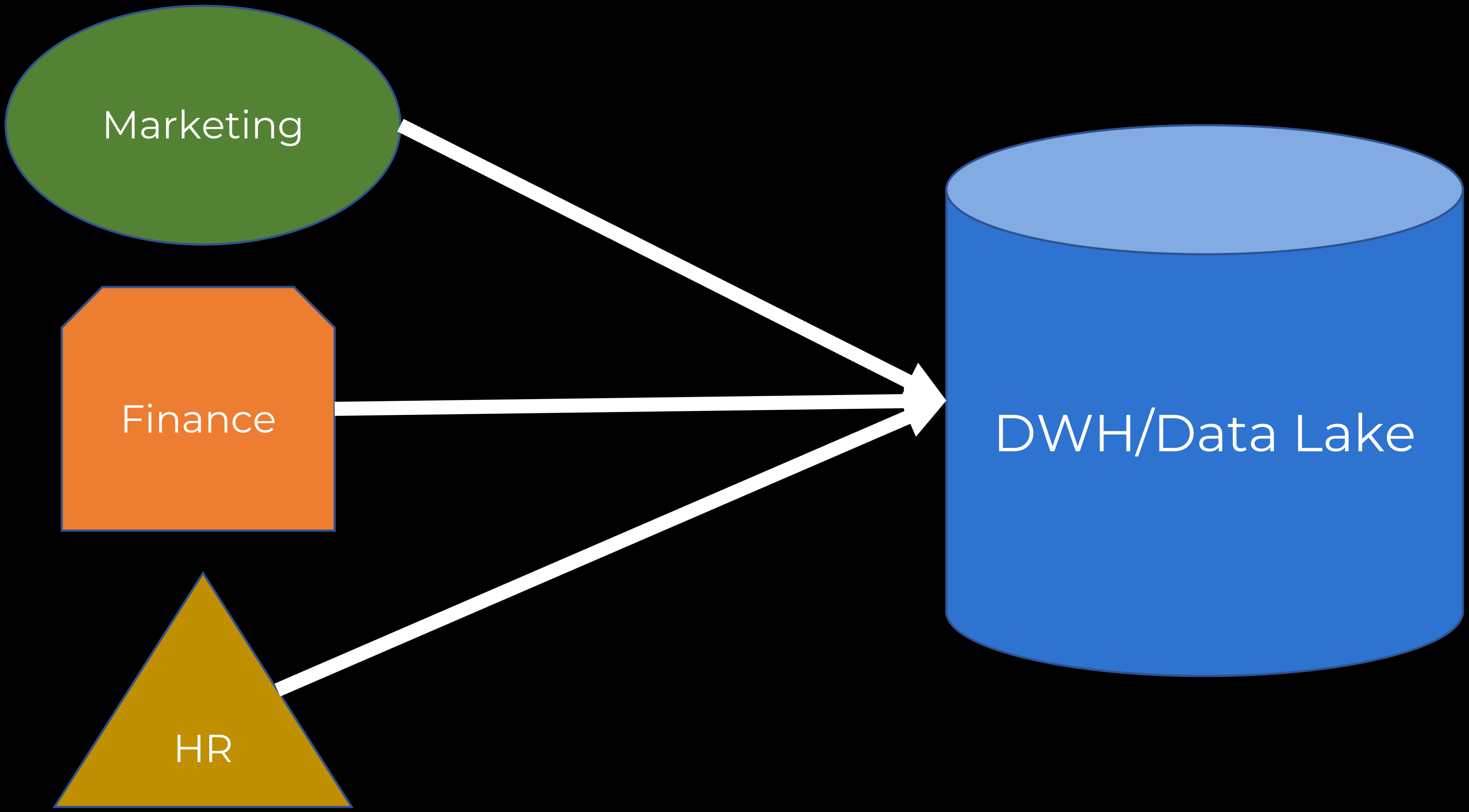
- Lack of data ownership
- Lack of data quality
- Organisational scaling

James Serra, “Data Mesh Defined”

Data Mesh Principles

Domain-Oriented Decentralized Data Ownership and Architecture

Domain-Oriented Decentralized Data Ownership and Architecture

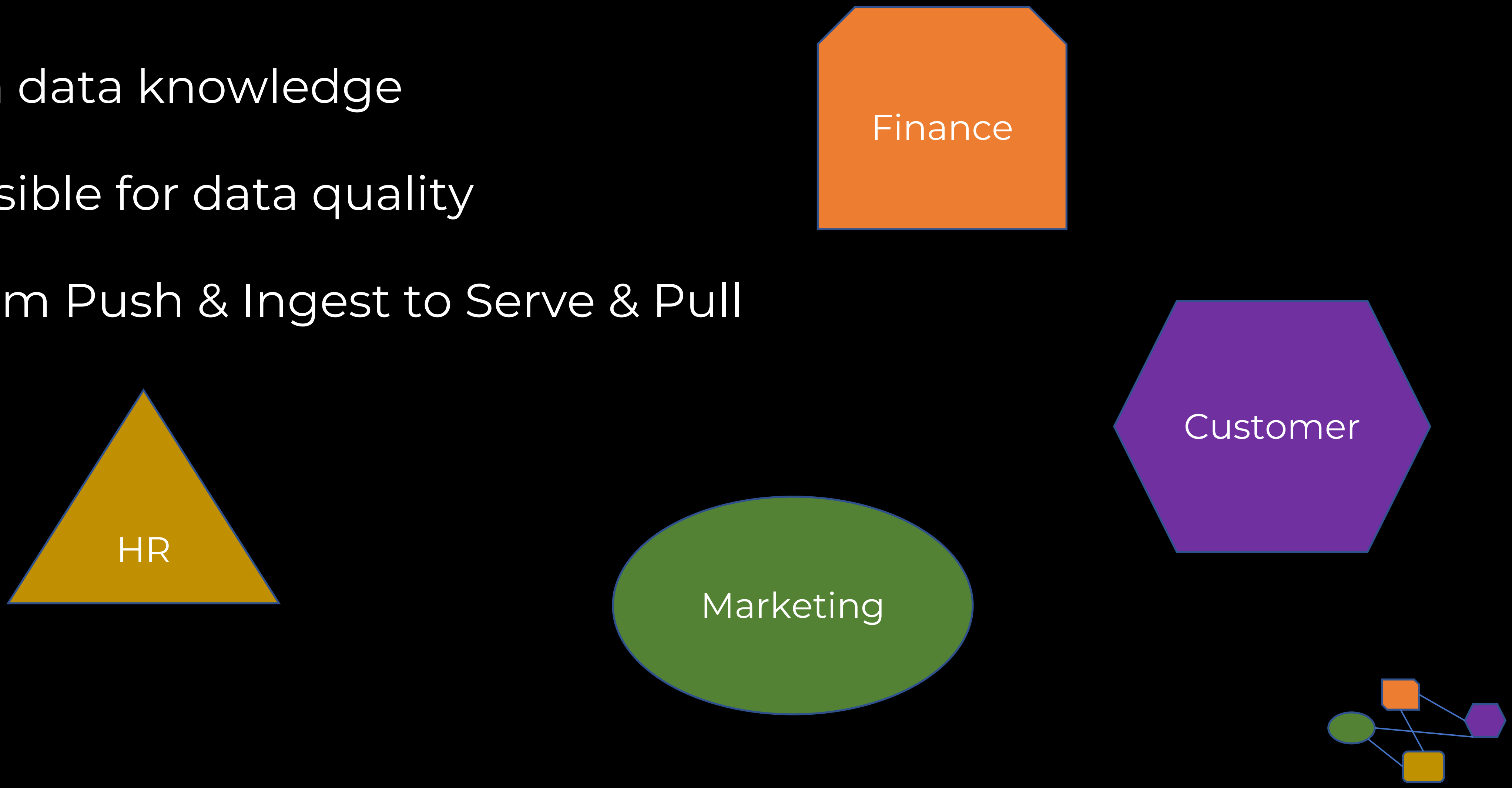


Domain-Oriented Decentralized Data Ownership and Architecture

Domain data knowledge

Responsible for data quality

Shift from Push & Ingest to Serve & Pull

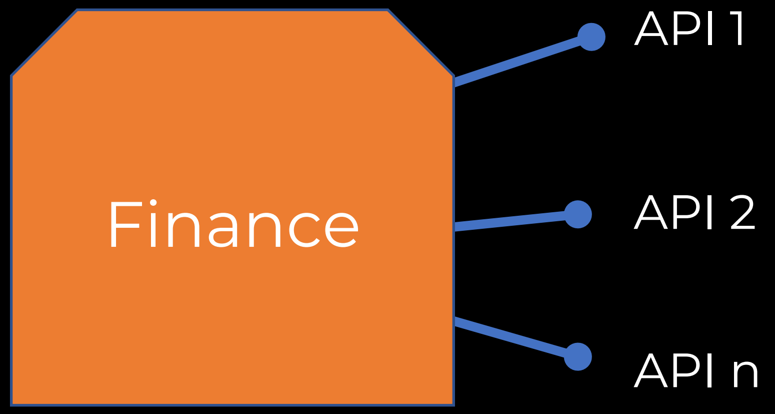


Data Mesh Principles

Domain-Oriented Decentralized Data Ownership and Architecture

Data as a Product

Data as a Product



Finance

API 1

API 2

API n

- Discoverable
- Adressable
- Trustworthy
- Self-describing
- Interoperable
- Secure

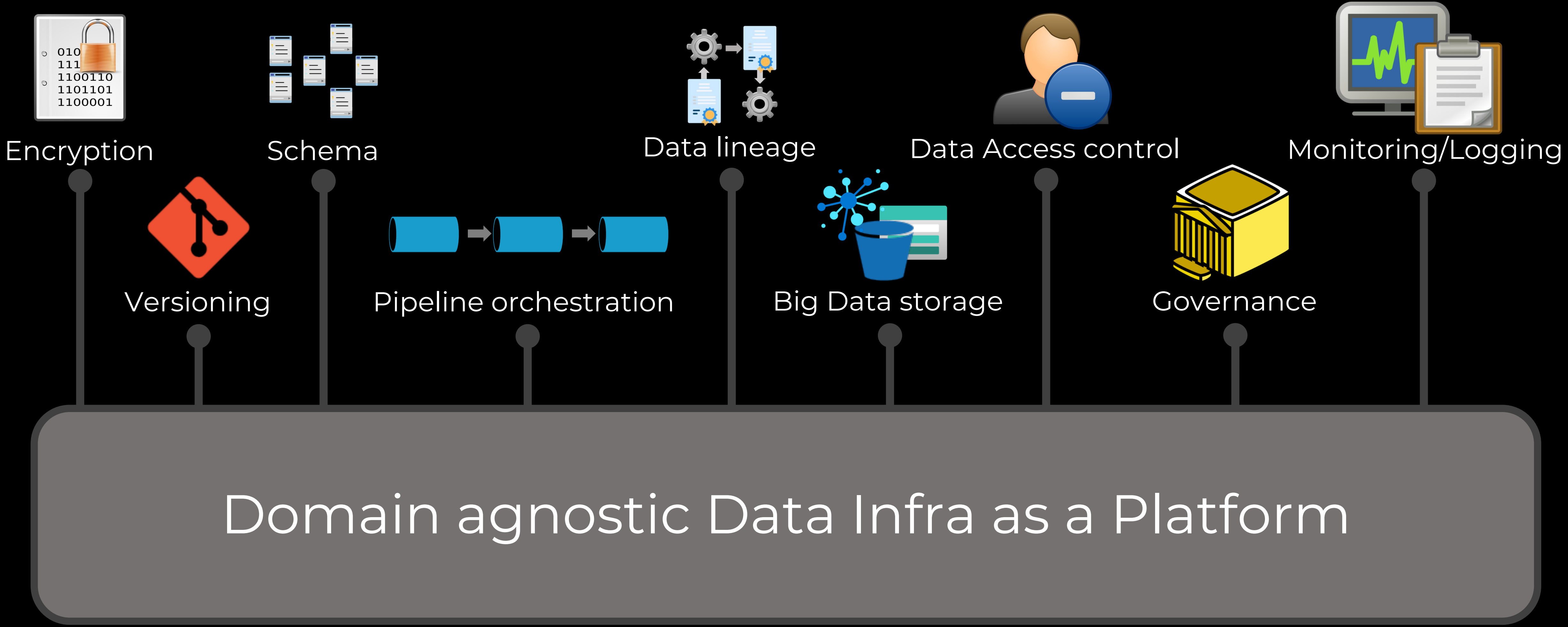
Data Mesh Principles

Domain-Oriented Decentralized Data Ownership and Architecture

Data as a Product

Self-Serve Data Infrastructure as a Platform

Self-Serve Data Infrastructure as a Platform



Encryption

Schema

Data lineage

Data Access control

Monitoring/Logging

Versioning

Pipeline orchestration

Big Data storage

Governance

Domain agnostic Data Infra as a Platform

Success criteria: Lowering lead time to create a new data product

Data Mesh Principles

Domain-Oriented Decentralized Data Ownership and Architecture

Data as a Product

Self-Serve Data Infrastructure as a Platform

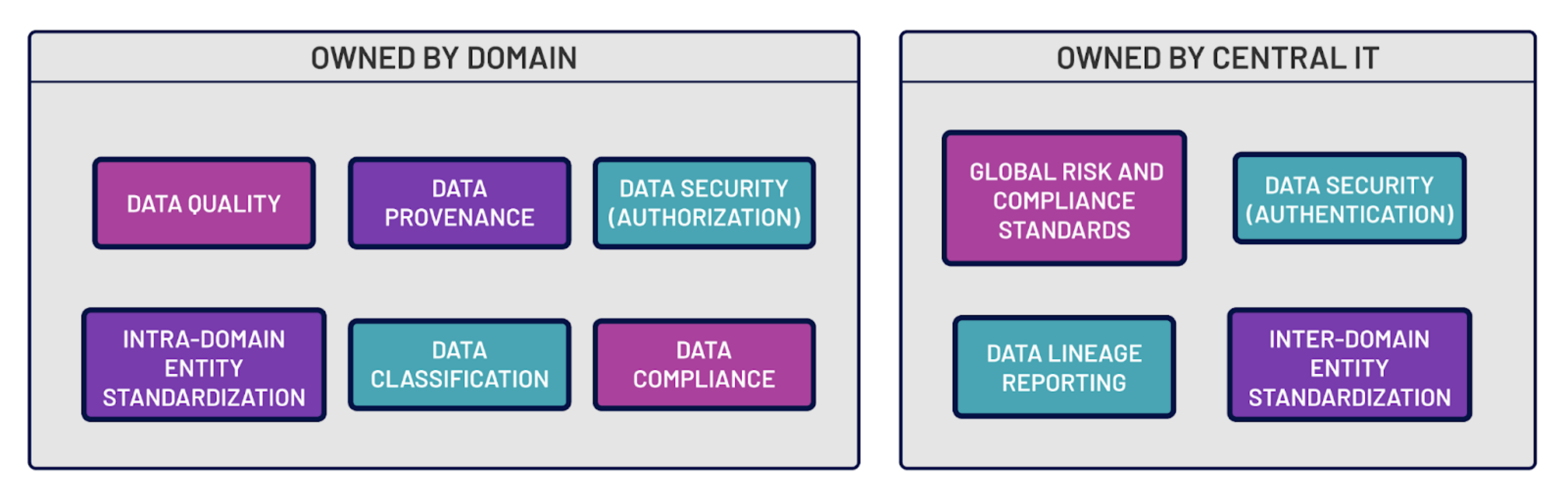
Federated Computational Governance

Federated computational governance

Shared responsibility between domains and central IT

Focus on interoperability

Example:



OWNED BY DOMAIN

- DATA QUALITY
- DATA PROVENANCE
- DATA SECURITY (AUTHORIZATION)
- INTRA-DOMAIN ENTITY STANDARDIZATION
- DATA CLASSIFICATION
- DATA COMPLIANCE

OWNED BY CENTRAL IT

- GLOBAL RISK AND COMPLIANCE STANDARDS
- DATA SECURITY (AUTHENTICATION)
- DATA LINEAGE REPORTING
- INTER-DOMAIN ENTITY STANDARDIZATION

https://blog.starburst.io/data-mesh-federated-computational-governance

Data Mesh paradigm shift

Centralized data platform → Ecosystem of data products

Extracting and loading → Discovering and using

Ingesting → Serving

Flowing data around → Publishing events as streams

Rune Ovlien Rakeie

“Data Mesh is the best thing invented since sliced bread”

Johan Ludvig Brattås

“Eh… 😬”

What if …

The Data Mesh concept is beating down open doors?

The principles of the Business Data Lake

Defined by Capgemini in 2013

1. Land all the information you can *as is with no modification*
2. Encourage LOB to create point solutions
3. Let LOB decide on the cost/performance for their problem
4. Concentrate governance on the critical points only
5. Consider the corporate view to be just another LOB view
6. Unstructured information is still information
7. Never assume the lake contains everything
8. Scale is driven by demands – scale down as well as up

Data Lakehouse

Defined by Databricks in 2019

Features of a data lakehouse:

- Transaction support
- Schema enforcement and governance
- BI support
- Storage is decoupled from compute
- Openness
- Support for diverse data types
- Support for diverse workloads
- End-to-end streaming
- Layered architecture
- Supports domain-oriented approach

However...

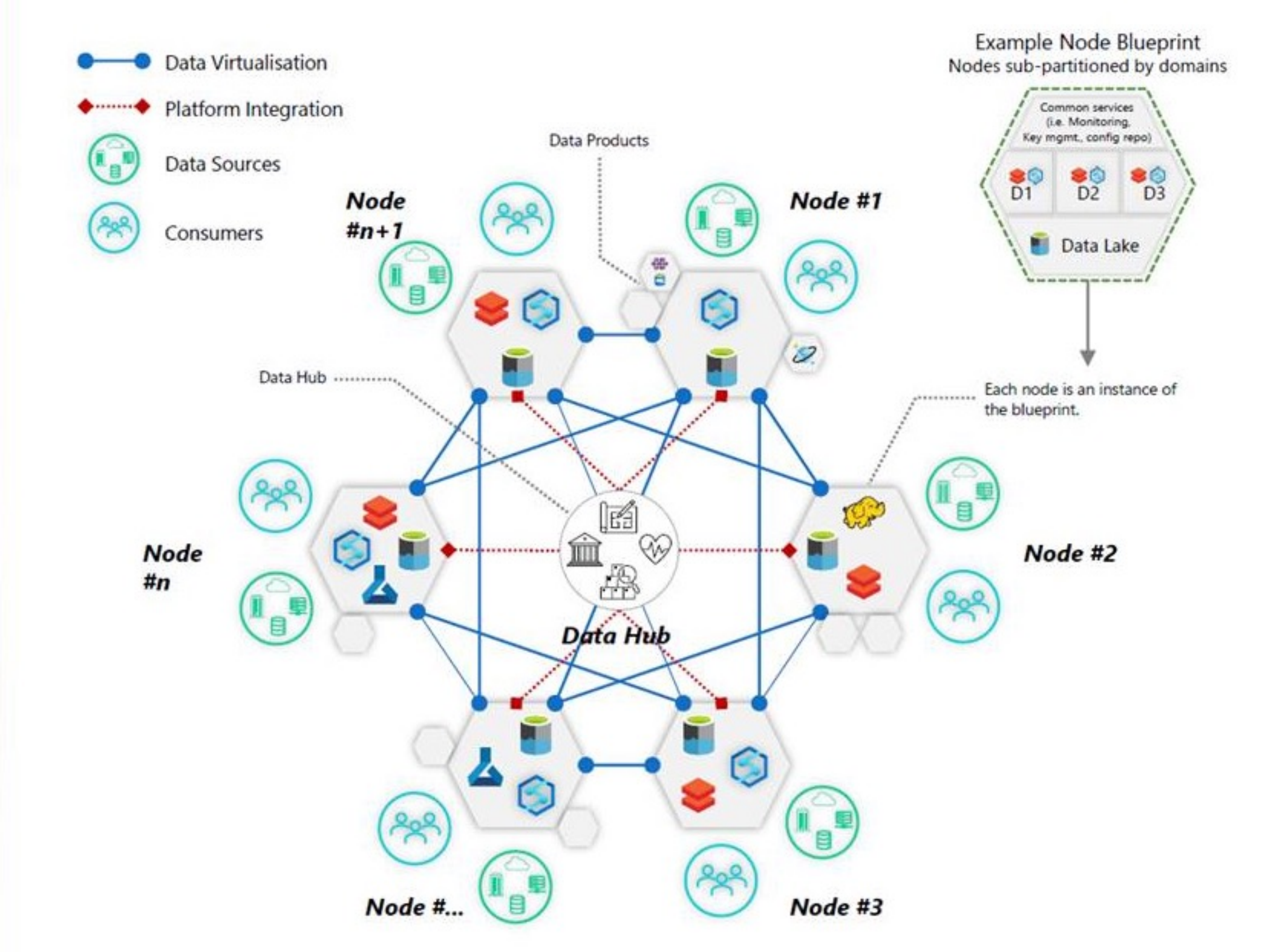
Data Mesh is less about technology

More about people and practices.

How to build a data mesh

- Data Mesh isn’t about tech
- However – it could look something like this…

Harmonised mesh



Cloud-scale analytics

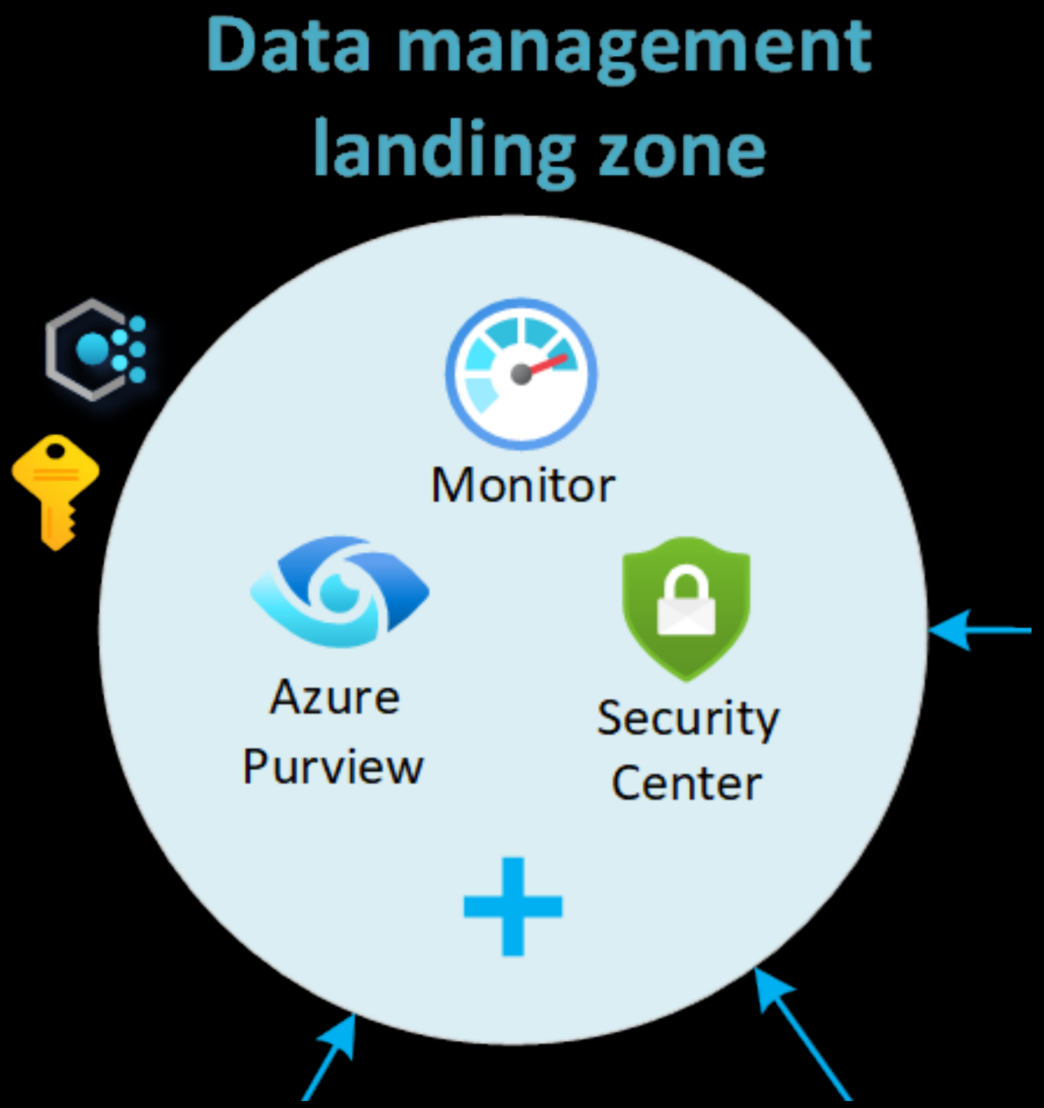
Cloud-scale is targeting data mesh-like patterns.

5 modules

- Data Management Zone
- Data Landing Zone(s)
- Data Integration – batch
- Data Integration – stream
- Data Product – Analytics & Data Science

- Your central data hub

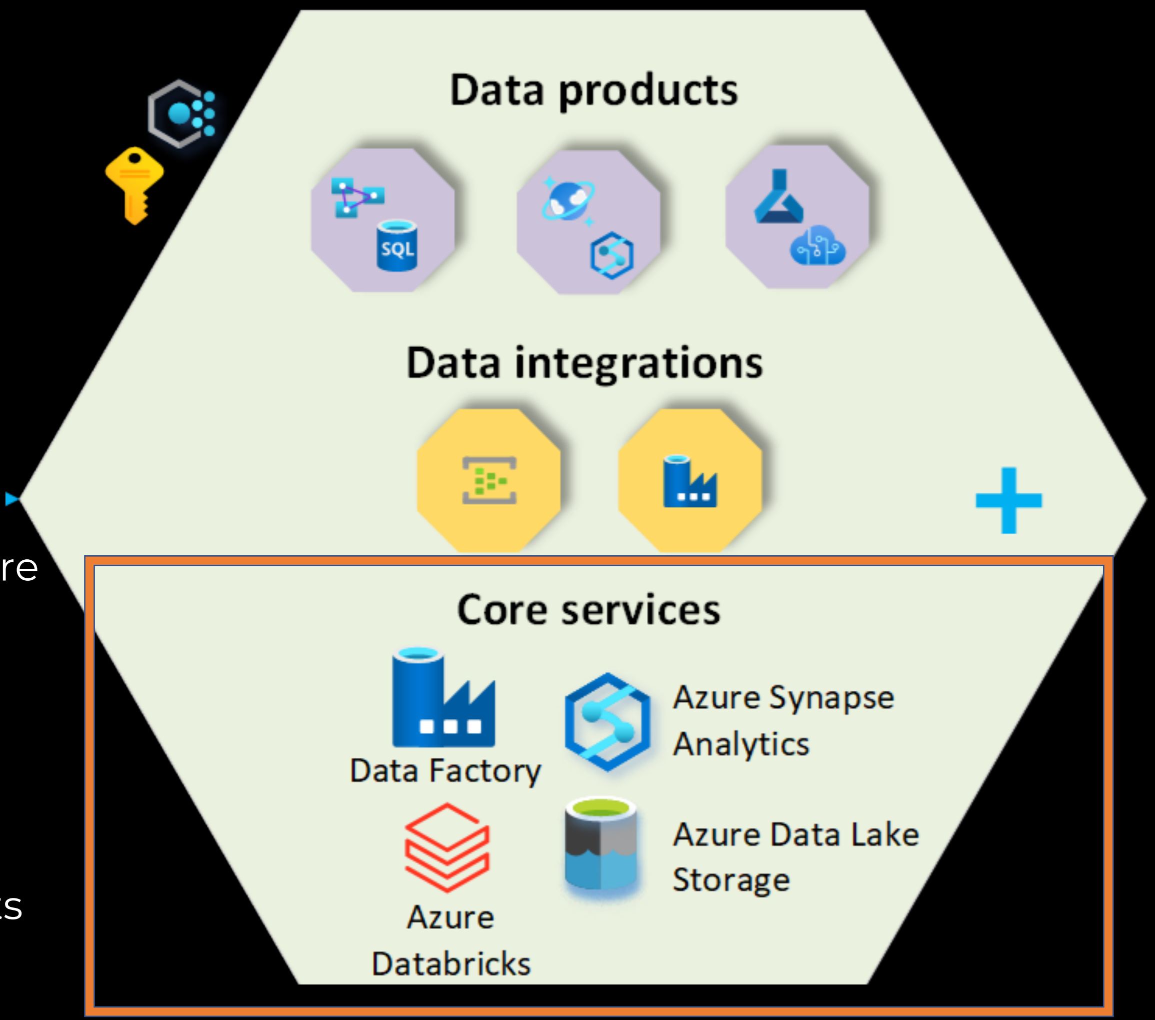
- Networking
- Central monitoring and security
- API Management
- Microservices hosting (AKS + CS)
- Purview
- Synapse Link
- Power BI link



Data Landing Zone

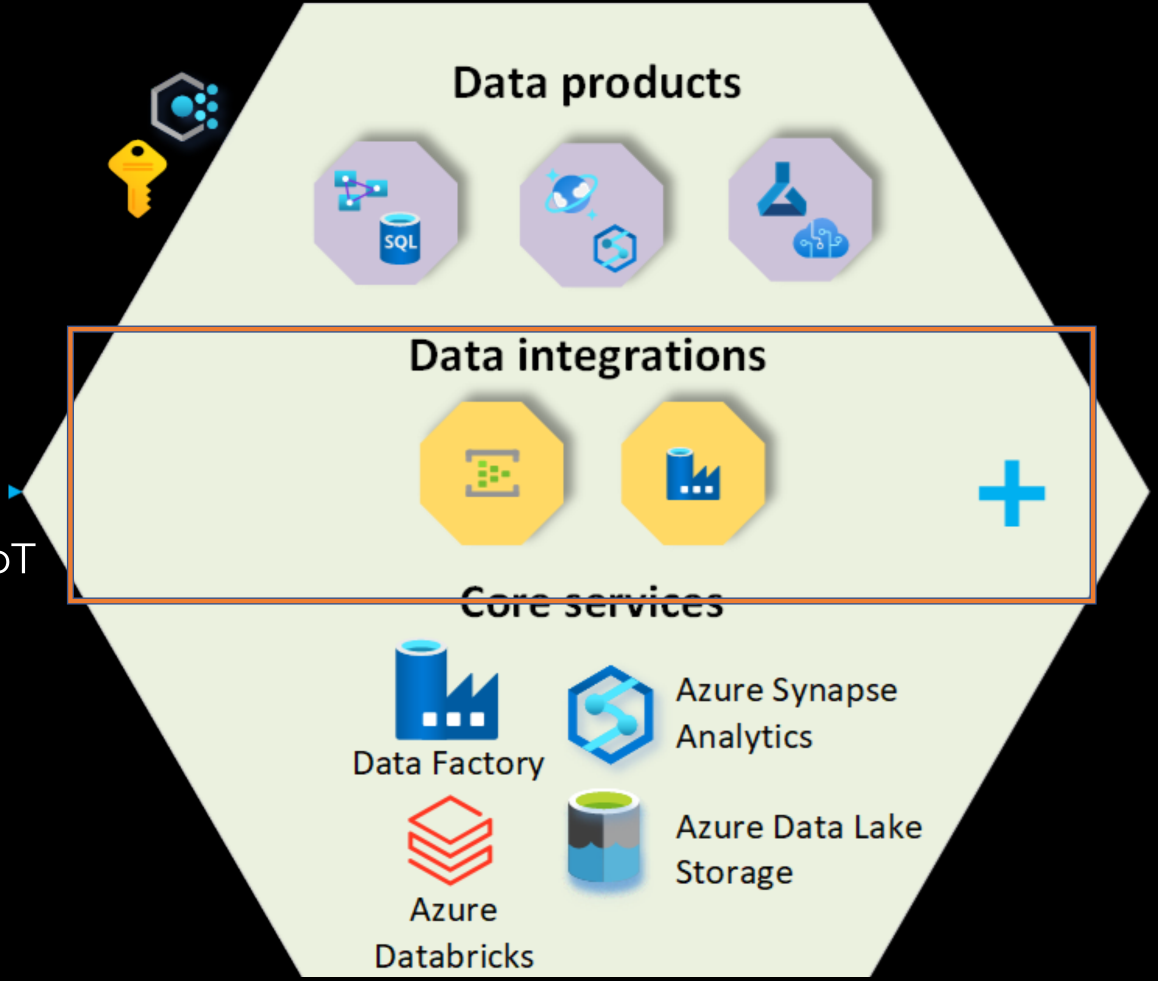
- As many as you need
- This is where your data is persisted – and workloads are executed
- Storage services
- Ingestion services
- Management services (networking, monitoring…)

- Also hosts your Integration services and data products



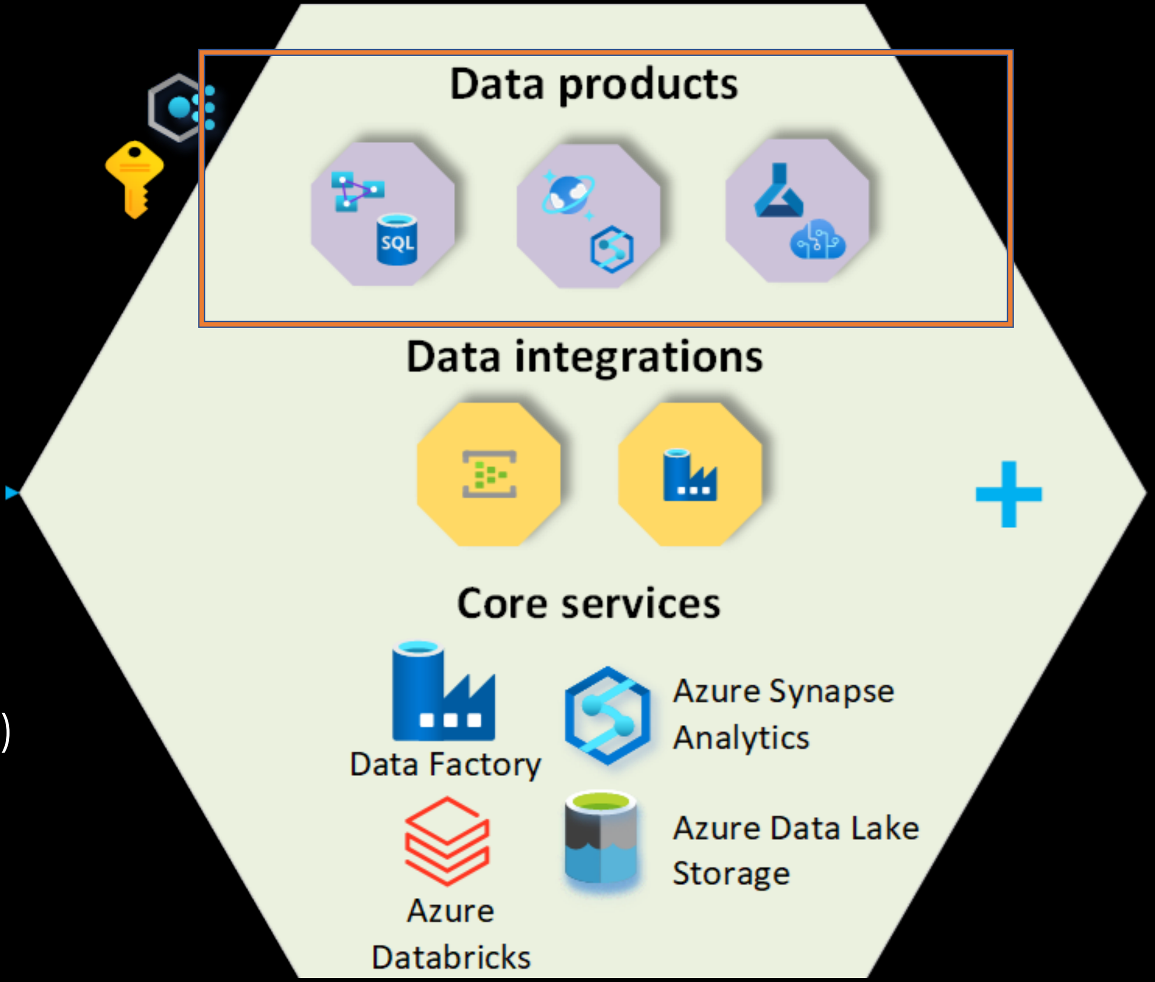
Data Integrations (batch & stream)

- Storage services
- Data orchestration – Data Factory & Event hub /IoT Hub
- Transformation – Data Factory / Synapse / Databricks / Stream Analytics
- Shared runtime services
- Metastores

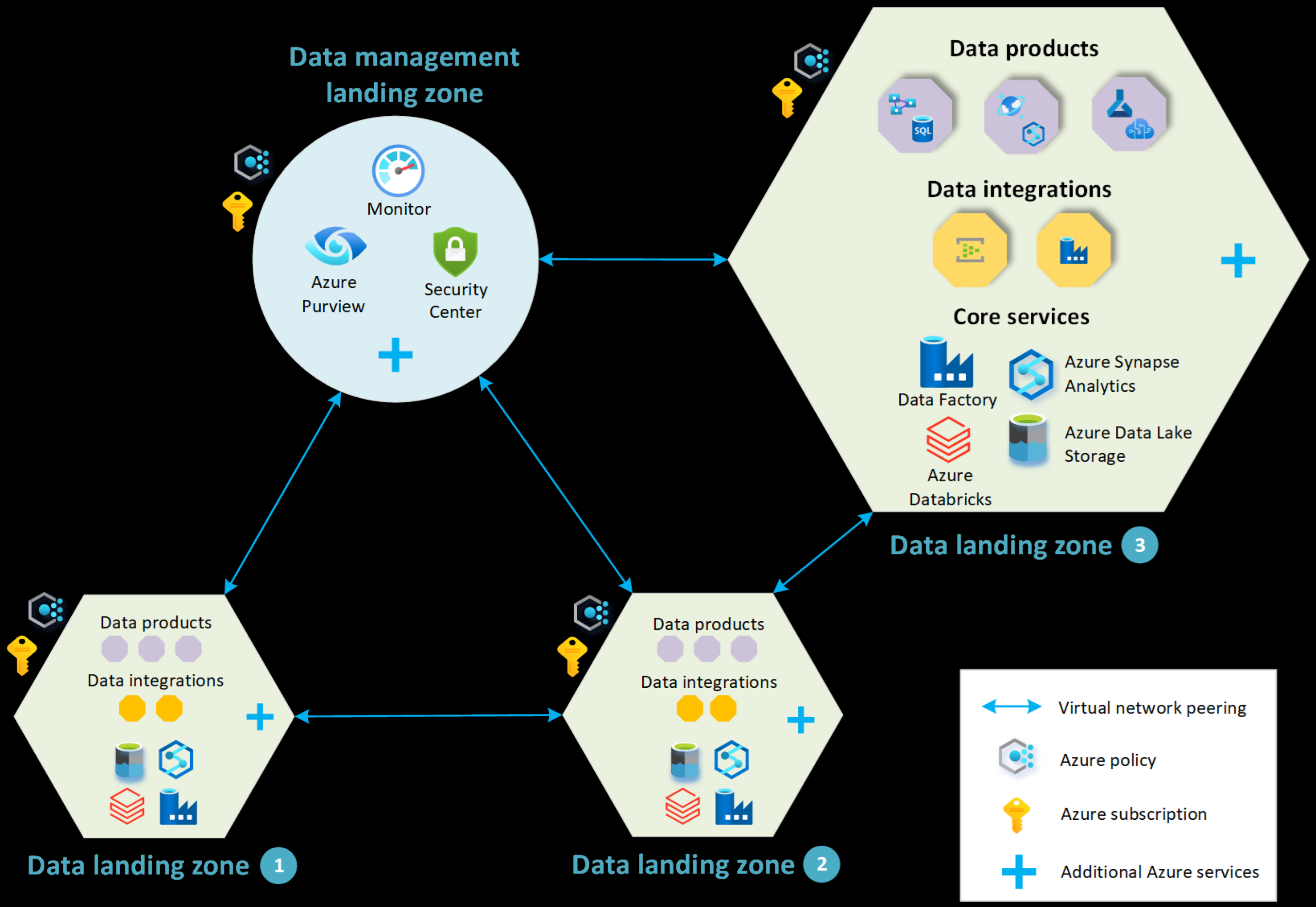


Data Product – Analytics & Data Science

- Synapse Workspace
- Databricks Workspace
- Key vaults
- Data Science services (Azure ML, Cognitive svc ++)
- Storage account



Cloud-Scale Analytics



Domains can be split in two

- Source Domain
- Consumer Domain

Not decentralized architecture.

Decentralized organization

Source-side challenges

- The producer(s) of data products need to follow principles and guidelines. But, more importantly, they need to understand that the data product they are producing will be consumed.

- On the source data domain side, the challenge is to make developers understand the analytical purpose of data products and adopting a data engineer mindset.

- One way to assist with this, is to have a central CoE team that can assist producers analyze their data and create contracts that they can follow – that adheres to the central principles.

Consumer-side challenges

- On the consumer data domain side, a core challenge can be to have business users and analysts understand their role. Not only consumers, but also producers and potentially owners of data products.

- Having analysts gain a data engineer mindset and taking ownership of the products they build, is usually an organizational challenge.

- Maintaining these data products over time can be another challenge. In particular, when consumer data products are built «on top of» other consumer data products, that are in turn built on top of source data products. Thus, forming a chain of data products, becoming more and more refined and purpose specific by each «evolution», but also dependent on previous "evolvement".

Contracts

Two types of contracts

Data Product specifications

- Content adhering to principles
- SLA

Data Product License Agreement

- Update intervals
- Expected data content and quality
- right to use for new products
- price – if any

Is this for everyone?

No

Most likely – 80% of organizations will never need this...

Though «everyone» wants to implement a data mesh

Considerations

- Do you have a complex source system landscape?
- Source systems mainly micro services?
- Lots of domains?
- Mature, data-driven organization?

Attollo

B3 Consulting Group

Quest Software

SQL Service

dbWatch

Transmokopter SQL

Catman Solution

DB24

Redgate Software

Rune Ovlien Rakeie

- Principal Cloud Architect Tietoevry

in
- /runeovlienrakeie

🐦
- @runeo34

Johan Ludvig Brattås

- Principal Solutions Architect Capgemini

in
- /johanludvig

🐦
- @intoleranse